Seeing Beyond the Scene: Enhancing Vision-Language Models with Interactional Reasoning

Anonymous ACL submission

Abstract

While scene graphs provide structured representations for multimodal reasoning, existing approaches predominantly encode static spatial relationships and lack role-aware interaction modeling, leaving VLMs unable to resolve complex agent-patient dependencies. Moreover, current fine-tuning paradigms fail to exploit the full structural potential of scene graphs and provide no effective self-feedback mechanism for iterative reasoning improvement. We propose ISGR, an interaction-augmented scene graph reasoning framework for vision-language models. ISGR first converts dense spatial graphs into query-conditioned interaction graphs through a Summarize-and-Align procedure guided by Interactional Chain-of-Thought. Secondly, it generates role-grounded instruction data and optimizes the model with a structure-aware reinforcement objective that rewards role consistency, entity disambiguation, and visual grounding while penalizing attention drift. Finally, a filtered self-refinement loop further improves pseudo graph quality without additional human annotation. Experiments on seven interaction-relevant benchmarks covering spatial reasoning, scene understanding, and multi-object reasoning show consistent gains over strong general VLM and scene-graph-enhanced baselines, with larger improvements on interaction-heavy reasoning tasks.

1 Introduction

Scene graphs (Yang et al., 2025; de Oliveira Souza et al., 2023) have emerged as a powerful structured representation for multimodal reasoning, underpinning a wide range of tasks including image captioning, visual grounding, and visual question answering (VQA). Modern vision-language datasets (Huang et al., 2024; Wang et al., 2025a) increasingly annotate diverse relationship types spanning both static spatial configurations (e.g., *"on"*, *"behind"*) and dynamic interactions (e.g., *"catching"*, *"watching"*). Despite this richness, existing scene graph methods (Liu et al., 2023c) usually encode relationships as flat triplets, treating interactions and spatial relations in the same way. This structural conflation obscures the compositional nature of visual interactions, where accurately understanding a scene requires not only detecting *what* relationships exist, but also identifying *who* acts as the agent and *who* serves as the patient within each interaction.

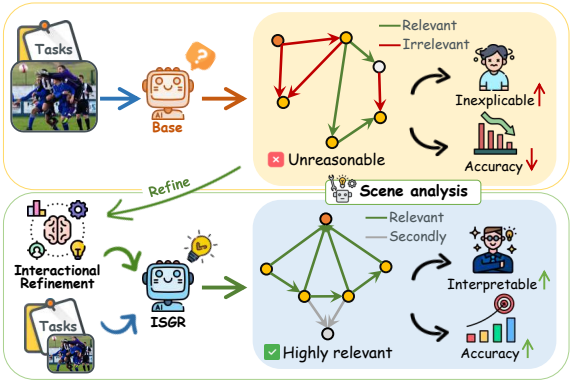

Figure 1: **Comparison with Traditional Scene Graph Methods.** Our interactional refinement process transforms flat, cluttered relationship triplets into structured, hierarchical interaction graphs, enabling accurate compositional reasoning over complex relational dynamics.

As illustrated in Figure 1, this flat representation becomes particularly problematic in complex multi-object scenarios. When multiple players engage in overlapping interactions, methods relying on flat scene graphs (Li et al., 2024a; Zheng et al., 2024) accumulate irrelevant relationships without structural differentiation, failing to answer compositional queries such as *"who will catch the ball?"* because they cannot distinguish active agents (e.g., a goalkeeper reaching for the ball) from background entities (e.g., surrounding defenders). Such role confusion critically undermines causal reasoning, where accurate inference requires explicit subject-object decomposition.

In this paper, we introduce **Interaction-augmented Scene Graph Reasoning (ISGR)**, a framework that replaces flat relationship triplets with structured, role-aware interaction representations to enhance compositional reasoning in VLMs. The core insight is that interaction modeling and model learning should co-evolve: better graphs improve reasoning, while stronger reasoning enables more accurate graph construction. To this end, ISGR integrates three coupled mechanisms. First, *Interactional Chain-of-Thought* (ICoT) prompting guides VLMs to decompose visual scenes into hierarchical interaction graphs with explicit subject-object role assignments, going beyond free-form chain-of-thought reasoning. Second, a structure-aware reinforcement strategy augments supervised fine-tuning via a customized GRPO framework (Shao et al., 2024), providing step-level feedback on role consistency, compositional validity, and visual grounding rather than evaluating holistic response quality. Third, a progressive self-refinement cycle lets the model iteratively re-annotate its training data with higher-quality graphs, creating a virtuous loop between representation quality and reasoning capability without additional human annotation.

We conduct extensive experiments on compositional VQA, visual grounding, and scene graph generation, demonstrating consistent and substantial improvements over existing methods. Our contributions are three-fold:

(1) We propose ISGR, an offline interaction-graph construction and training framework that converts flat spatial scene graphs into query-conditioned, role-aware interaction graphs, explicitly modeling agent-action-patient structures.

(2) We introduce Interactional Chain-of-Thought (ICoT), a structured graph transformation procedure that progressively abstracts dense spatial relations into salient interaction graphs and generates role-grounded instruction data.

(3) We design a structure-aware interaction reinforcement objective with decomposed rewards for role consistency, entity disambiguation, visual grounding, and attention-drift suppression, together with a filtered self-refinement loop for improving pseudo graph quality.

2 Related Work

VLMs for Scene Understanding. Recent VLMs (Li et al., 2024a; Liu et al., 2024; Bai et al., 2023) have advanced scene understanding through improved visual encoders, spatial reasoning, and language supervision, with extensions to video, 3D, and embodied settings (Xiao et al., 2025). Further work addresses open-vocabulary segmentation (Yuan et al., 2024a; Du et al., 2024), zero-shot counting (Zhai et al., 2025), and domain adaptation (Zhou and Zhou, 2024). Despite these advances, contemporary VLMs represent interactions as flat, unstructured predicates, limiting role-dependent reasoning in complex scenes (Wang et al., 2025b).

Scene Graphs for Visual Reasoning. Scene graphs have proven effective for topological representation (Wu et al., 2021), robotic navigation (Greve et al., 2024; Zhu et al., 2021), and temporal relational modeling (Wu et al., 2025a; Feng et al., 2025). VLM-integrated approaches inject graph encoders (Huang et al., 2024; Wang et al., 2025a) or linearize graph cues as prompts (Zheng et al., 2024), while grounding and relation alignment methods (Liu et al., 2023c) improve structured prediction but remain confined to closed predicate spaces.

Self-Improving LLMs. Recent studies have explored training-time self-improvement, where LLMs refine their own responses or preferences through self-generated feedback. Self-rewarding models use the LLM itself to provide reward signals for iterative alignment (Yuan et al., 2024b), while self-play fine-tuning improves models by contrasting responses from different policy iterations (Chen et al., 2024). Meta-rewarding further enhances this process by letting the model evaluate its own judgments (Wu et al., 2025b). Different from these text-centric methods, ISGR performs self-refinement over structured multimodal representations, where feedback is tied to role-aware interaction graphs, visual grounding, and entity disambiguation.

3 Interaction-augmented Scene Graph Reasoning

Contemporary vision-language models exhibit remarkable capabilities in object recognition and spatial reasoning, yet they frequently falter when confronted with complex interactional dynamics within visual scenes. This limitation stems from their predominant focus on static spatial relationships while neglecting the temporal and functional interactions that define meaningful scene under-

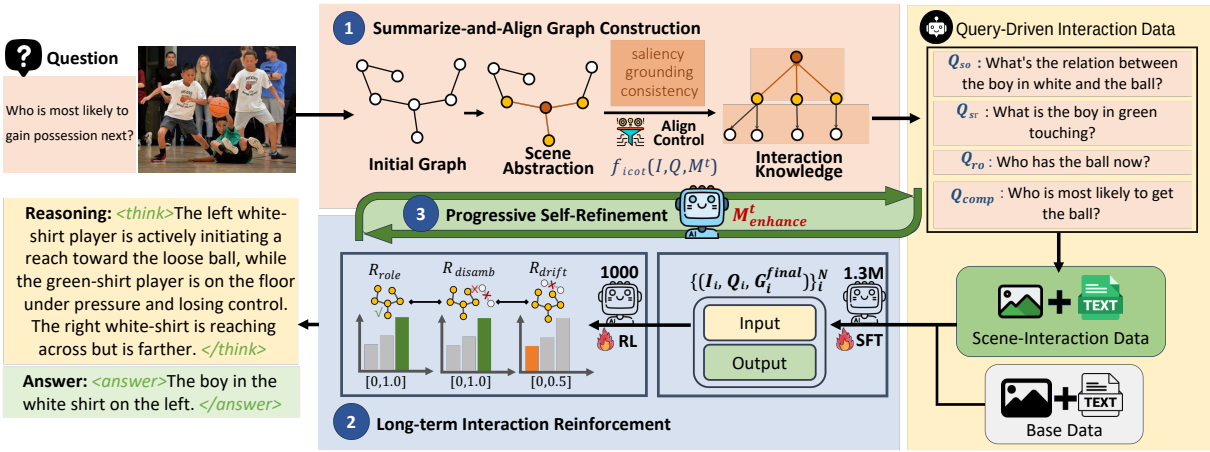

Figure 2: **Overview of ISGR.** (1) Summarize-and-Align Graph Construction constructs interaction graphs and derives role-annotated QA data; (2)Long-term Interaction Reinforcement applies dual-phase learning with a decomposed reward (R_{role}, R_{disamb}, R_{drift}) to internalize structural reasoning; (3)Progressive Self-Refinement iteratively improves both graph quality and model capability without additional human annotation.

standing. To address this gap, we introduce **ISGR (Interaction-augmented Scene Graph Reasoning)**, a framework that enhances VLMs with structured, role-aware interaction representations. As illustrated in Figure 2, ISGR consists of four stages: Summarize-and-Align graph construction, long-term interaction reinforcement, and progressive self-refinement.

The architectural philosophy underlying ISGR rests on a critical observation: effective interaction reasoning requires not merely the identification of objects and their spatial arrangements, but rather the dynamic interplay between comprehensive scene understanding and accumulated experiential knowledge. Our framework materializes this insight through two interconnected modules that operate in a mutually reinforcing cycle, where enhanced scene construction capabilities inform memory development, while refined memory systems enable more sophisticated scene analysis.

The iterative refinement process can be formally expressed as:

$$\mathcal{T}^i = (\mathcal{G}^i, \mathcal{S}, \mathcal{M}^i, \mathcal{F}), \tag{1}$$

where the scene graph $\mathcal{G}^i$ evolves through iterations, guided by multimodal input $\mathcal{S} = \{Q, I\}$, processed by an increasingly sophisticated model $\mathcal{M}^i$, and refined through operation functions $\mathcal{F}$. The superscript i denotes the iterative evolution of both the scene graph and the model capability.

3.1 Summarize-and-Align Graph Construction

The foundation of interaction-aware scene understanding lies in constructing representations that capture both explicit spatial configurations and implicit functional relationships within visual scenes. As illustrated in Figure 3, our Summarize-and-Align Graph Construction module addresses this challenge through a progressive abstraction process, transforming raw visual information into structured, interaction-centric representations.

3.1.1 Multi-level Scene Abstraction

The construction begins with comprehensive scene parsing, establishing multiple abstraction levels to capture the hierarchical nature of visual understanding. The initial spatial graph construction identifies fundamental scene elements:

$$\mathcal{G}^s = \langle \mathcal{V}^s, \mathcal{E}^s \rangle = f_{\text{spatial}}(I, \mathcal{M}^{base}). \tag{2}$$

Here, $\mathcal{V}^s$ includes all detectable entities, while $\mathcal{E}^s$ captures their spatial relationships. The function f_{spatial} utilizes the base model $\mathcal{M}^{base}$ to perform this initial parsing, creating a foundational layer for further reasoning. We implement f_{spatial} through prompted inference with the multimodal LLM:

> Generate a spatial scene graph identifying objects and their spatial relationships for {*image*}

Building on this foundation, a contextual abstraction mechanism filters and prioritizes scene

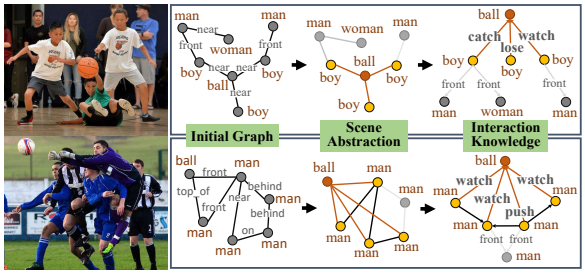

Figure 3: **Summarize-and-Align Graph Construction.** Progressive scene graph transformation through our Summarize-and-Align Graph Construction. The Initial Graph contains dense spatial relationships, Scene Abstraction filters irrelevant connections, and Interaction Knowledge identifies role-aware interactions (e.g., "*man catch ball*", "*man watch*") with hierarchical structures.

elements based on their relevance to the query:

$$\mathcal{G}^a = f_{\text{abstract}}(\mathcal{G}^s, I, Q, \mathcal{M}^{base}). \quad (3)$$

This process operates under three complementary constraints: the Drift constraint emphasizes salient entities, the Role constraint filters based on task pertinence, and the disambiguation constraint resolves ambiguities in entity references. This process can be implemented as:

> Create an abstract graph of {*spatial scene graph*} that focuses only on elements relevant to the core scene: {*image*}.

3.1.2 Interaction-Centric Alignment

The key innovation lies in the alignment phase, where we convert the abstracted spatial representation into an interaction-centric graph that captures dynamic relationships between scene elements, guided by our Interactional Chain-of-Thoughts (ICoT) methodology:

$$\mathcal{G}^{inter} = f_{\text{ICoT}}(\mathcal{G}^a, I, Q, \mathcal{M}^{base}). \quad (4)$$

ICoT systematically identifies and formalizes interactions through a structured reasoning chain. Subject identification establishes primary actors, action recognition determines behaviors, object identification specifies targets, and relation formalization converts observations into structured triplets. This is implemented through:

> Using the {*abstract graph*}, identify all interactions between entities that are relevant to the core scene: {*image*}. For each interaction, specify the subject, action, and object.

This alignment culminates in a final abstraction step that synthesizes interaction knowledge into a coherent, query-relevant representation:

$$\mathcal{G}^{final} = f_{\text{align}}(\mathcal{G}^{inter}, I, Q, \mathcal{M}^{base}). \quad (5)$$

The alignment function f_{align} ensures saliency (emphasizing significant interactions), grounding (maintaining visual correspondence), and consistency (preserving logical coherence). An exemplar implementation is as follows:

> Using the {*interaction knowledge*}, further abstract the scene by identifying the most relevant interactions for the core scene: {*image*}. Please specify the subject, action, and object while ensuring adherence to the saliency, grounding, and consistency constraints.

To support comprehensive interaction reasoning, our framework accommodates four query types that extract relational information from the **query-driven interaction**: Object-Object Queries $Q_{oo}(o_1, o_2)$, Subject-Relation Queries $Q_{sr}(s, r)$, Relation-Object Queries $Q_{ro}(r, o)$, and Comprehensive Queries $Q_{\text{cs}}(e)$. These query mechanisms collectively enable the generation of training instances that capture the full spectrum of interaction reasoning scenarios.

3.2 Long-term Interaction Reinforcement

While the Summarize-and-Align module provides sophisticated scene analysis capabilities, the true power of interaction reasoning emerges through the accumulation and application of experiential knowledge. Our Long-term Interaction Reinforcement (LIR) module addresses this requirement by developing persistent memory mechanisms that enhance the model's ability to recognize, interpret, and reason about complex interactional patterns.

3.2.1 Structure-aware Model Development

The LIR process transforms the base model into an interaction-aware system through targeted enhancement. This transformation integrates traditional visual-linguistic training data $\mathcal{D}^{base}$ with our constructed interaction-centric dataset $\mathcal{D}^{interact}$,

where:

$$\mathcal{D}^{inter} = \{(I_i, Q_i, \mathcal{G}_i^{final})\}_{i=1}^{N}. \quad (6)$$

Each training instance comprises an image I_i, a contextually relevant query Q_i, and the corresponding interaction-augmented scene graph $\mathcal{G}_i^{final}$ generated through our Summarize-and-Align process.

3.2.2 Dual-Phase Memory Consolidation

The supervised fine-tuning phase establishes foundational interaction recognition by minimizing the negative log-likelihood over the interaction-augmented graph sequences:

$$\mathcal{L}_{\text{SFT}} = -\sum_{t=1}^{T} \log P_{\mathcal{M}}(g_t \mid I, Q, g_{<t}), \quad (7)$$

where g_t denotes the t-th token of the target interaction graph $\mathcal{G}$, $g_{<t}$ denotes all preceding tokens, and T is the total sequence length. The model $\mathcal{M}$ is thus trained to autoregressively generate structured, role-annotated interaction graphs conditioned on the input image I and query Q.

The reinforcement phase refines interaction reasoning via GRPO (Shao et al., 2024) with a decomposed reward that supervises three structural dimensions of the generated response y:

$$\begin{aligned} R(y_k) =& \lambda_1 \cdot \mathcal{F}_{role}(y_k) + \lambda_2 \cdot \mathcal{F}_{disamb}(y_k) \\ &- \lambda_3 \cdot \mathcal{F}_{drift}(y_k), \end{aligned} \quad (8)$$

where $\mathcal{F}_{\text{role}}$ rewards correct agent-patient role assignments, $\mathcal{F}_{\text{disamb}}$ rewards unambiguous subject-object distinctions in multi-object scenarios, and $\mathcal{F}_{\text{drift}}$ penalizes attention drift toward irrelevant entities. The policy is then optimized by:

$$\mathcal{L}_{\text{IRR}} = -\sum_{k=1}^{K} \hat{A}_k \log P_{\mathcal{M}}(g_k \mid I, Q, g_{<k}), \quad (9)$$

where $\hat{A}_k$ is the advantage estimate computed from group-relative reward normalization across sampled responses. Full computation details are provided in the Appendix B.

3.3 Progressive Self-Refinement

Traditional training on fixed datasets limits models to their initial distribution. ISGR addresses this through an iterative self-refinement cycle, where improved interaction understanding progressively yields higher-quality training graphs.

Formally, at iteration t, the current model $\mathcal{M}_t^{\text{enh}}$ applies ICoT to construct an updated dataset:

$$\mathcal{D}_{t+1}^{\text{inter}} = \{(I_i, Q_i, \mathcal{G}_{i,t+1})\}_{i=1}^{N} \quad (10)$$

where $\mathcal{G}_{i,t+1}$ exhibits more accurate role assignments, clearer hierarchical structure, and better visual grounding than $\mathcal{G}_{i,t}$, since $\mathcal{M}_t^{\text{enh}}$ has internalized compositional patterns from the previous iteration. Dual-phase learning on $\mathcal{D}_{t+1}^{\text{inter}}$ then produces $\mathcal{M}_{t+1}^{\text{enh}}$, continuing the cycle. The self-refinement process improves performance in early iterations but may introduce noise in later rounds. Therefore, we adopt validation-based early stopping and reward-based filtering to prevent pseudo-label drift. Detailed filtering rules and update criteria are provided in Appendix C.

4 Experiments

Our experimental evaluation validates ISGR's memory-centric approach by systematically assessing how Long-term Interaction Reinforcement transforms interaction understanding. We address three fundamental questions: (1) whether interaction-augmented scene graphs improve over spatial-only reasoning; (2) how Long-term Interaction Reinforcement affects reasoning performance; and (3) whether progressive self-refinement improves the model across iterations.

4.1 Experimental Setup

Dataset Construction. We build our dataset through the iterative refinement pipeline described in Section 2.3. The base data integrates LLaVA-v1.5-mixed-665k (Liu et al., 2024) for general vision-language capabilities, OpenImages (Kuznetsova et al., 2020) (176K images) for spatial relationship grounding, LVIS-Instruct-4V (Wang et al., 2023) for fine-grained recognition, and LRV-Instruct (Liu et al., 2023b) for hallucination reduction. We apply ICoT to transform 300K instances into interaction-augmented scene graphs with explicit role annotations. We construct a small-scale set of 841K instances by combining LLaVA-v1.5-mixed-665K with 176K interaction-augmented OpenImages instances. For the medium-scale set, after filtering and deduplication, we obtain 1,371K instances from all sources. We additionally curate 1k high-quality compositional reasoning instructions for the structure-aware reinforcement learning phase, each requiring multi-hop reasoning over interaction chains.

Table 1: Performance comparison across benchmarks. Data size refers to training sample count. We propose three models: ISGR-S (0.8M), ISGR-M (1.3M), and ISGR-IRR. Benchmarks (VizWiz, VSR, MMT-Bench, RealworldQA, A-Bench, SEEDBench, MME) assess various visual reasoning capabilities.

Model	Data Size	VizWiz	VSR	MMT	RWQA	A-Bench	SEEDB	MME
InstructBLIP(2023)	130M	33.4	52.1	36.7	37.5	45.6	46.8	1352.8
MMAlaya(2024)	-	-	-	38.6	45.2	-	58.8	1356.0
Kosmos2(2024)	>20M	-	-	25.5	32.7	33.2	-	721.1
MoE-LLaVA(2024)	2.2M	43.9	-	-	-	-	-	-
LLaVA-SG(2024)	1.2M	52.2	58.9	46.3	52.2	64.5	65.9	1401.8
SmoleVLM2(2025)	3.3M	-	-	44.7	49.9	63.7	-	1408.3
Emu3-Chat(2025)	13.5M	-	58.9	43.4	54.1	61.2	-	1410.5
SGVLM(2025)	1.2M	50.5	54.9	42.6	49.5	63.5	65.3	1378.5
LLaVA v1.5-7B-S	0.6M	45.9	54.1	40.9	47.5	61.3	60.0	1352.5
ISGR-S(**Ours**)	0.8M	49.5 (+3.6)	55.7 (+1.6)	44.1 (+3.2)	51.8 (+4.3)	63.8 (+2.5)	65.1 (+5.1)	1403.3 (+50.8)
LLaVA v1.5-7B-M	1.2M	49.6	54.5	41.1	48.2	62.1	60.7	1256.3
ISGR-M(**Ours**)	1.3M	51.3 (+1.7)	60.3 (+5.8)	45.2 (+4.1)	52.2 (+4.0)	64.7 (+2.6)	65.1 (+4.4)	1291.7 (+35.4)
LLaVA-IRR	1.2M+1K	50.8	55.6	43.2	51.7	63.4	63.5	1344.1
ISGR-IRR(**Ours**)	1.3M+1K	**54.5** (+3.7)	**60.6** (+5.0)	**48.7** (+5.5)	**54.6** (+2.9)	**66.3** (+2.9)	**67.9** (+4.4)	**1414.1** (+70.0)

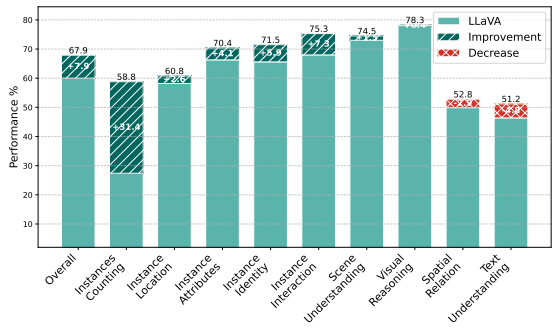

Figure 4: **Category Performance on SEEDBench.** ISGR-IRR shows significant improvements over the LLaVA-v1.5 baseline across most categories.

Table 2: Ablation study on Long-term Interaction Reinforcement showing the impact of different reward components (Role, Disambiguation, Drift) on ISGR-IRR model performance across benchmarks.

Rol.	Dis.	Dri.	MMT	A-Bench	SEEDB
✗	✗	✗	43.2	63.4	63.5
✓	✗	✗	48.6	65.4	67.2
✗	✓	✗	49.1	65.2	67.1
✗	✗	✓	48.6	65.4	66.7
✓	✓	✓	**48.7**	**66.3**	**67.9**

Implementation Details. We trained our models on 8 NVIDIA A100 GPUs using the LLaVA-v1.5 (7B) model. Supervised fine-tuning was conducted from pre-trained checkpoints with a per-device batch size of 8. Following fine-tuning on the interaction-augmented dataset, models underwent interaction reasoning reinforcement using 1k high-quality instruction examples. The reward function R was configured with $\lambda_1 = 0.4$, $\lambda_2 = 0.4$, and $\lambda_3 = 0.2$. We set t=2 for the final ISGR-IRR model. Figure 6 reports an additional diagnostic sweep over refinement steps from 0 to 5 to analyze the effect of repeated self-refinement.

4.2 Evaluation Benchmark

We selected benchmarks to comprehensively evaluate both general vision-language capabilities and specific interactional reasoning skills: VizWiz(Gurari et al., 2018)(real-world accessibility questions), VSR(Liu et al., 2023a)(visual spatial reasoning), RealWorldQA(x.ai, 2024)(practical spatial understanding), MMT-Bench(Ying et al., 2024)(recognition, localization, and reasoning), SEEDBench(Li et al., 2024b)(interaction, spatial and temporal understanding), and A-Bench(Zhang et al., 2024)(scene understanding).

4.3 Baseline Models

We compare against two categories of baselines. For general VLMs, we report the general reasoning results from VLMEvalKit (Duan et al., 2024), a unified evaluation toolkit for large multimodal models. These baselines include MMAlaya (Ltd., 2024), Emu3-Chat (Wang et al., 2024), InstructBLIP (Dai et al., 2023), Kosmos2 (Peng et al., 2024), SmolVLM2 (Marafioti et al., 2025), MoE-LLaVA (Lin et al., 2024), and LLaVA v1.5 (Liu et al., 2024). Scene-graph-enhanced models include SceneGraphVLM (Makarov et al., 2026) and LLaVA-SG (Wang et al., 2025a), whose results are reproduced based on their original implementations.

Our proposed models include: LLaVA-IRR (directly optimized with interaction reasoning reinforcement), ISGR (fine-tuned with interaction-augmented scene graph dataset in two versions: ISGR-S with 0.8M data and ISGR-M with 1.3M data), and ISGR-IRR (further optimization through interaction reasoning reinforcement).

4.4 Main Results

Table 1 shows that ISGR consistently improves multimodal reasoning performance. Compared with LLaVA-v1.5-7B-M, ISGR-IRR achieves clear gains on most benchmarks, including VSR (+6.1), MMT-Bench (+7.6), RealWorldQA (+6.4), A-Bench (+4.2), SEEDBench (+7.2), and MME (+157.8). The improvements are especially large on VSR, MMT-Bench, RealWorldQA, and SEED-Bench, indicating that ISGR is more effective on tasks requiring spatial relation understanding, object interaction reasoning, and compositional scene interpretation.

Compared with the strongest scene-graph-enhanced baseline LLaVA-SG, ISGR-IRR achieves consistent gains across all benchmarks, including VizWiz (+2.3), VSR (+1.7), MMT-Bench (+2.4), and MME (+12.3). These improvements span compositional question answering, spatial relation understanding, and general multimodal reasoning, indicating that explicitly modeling structure-aware interaction graph provides fundamentally richer reasoning cues than flat relational graph.

The comparison among our variants further verifies the effectiveness of each component. ISGR-S uses only 0.8M training samples, fewer than LLaVA-v1.5-7B-M with 1.2M samples, but still improves MMT-Bench (+3.0), RealWorldQA (+3.6), SEEDBench (+4.4), and MME (+147.0). After scaling to 1.3M samples, ISGR-M further improves most reasoning benchmarks, especially VSR (+4.6 over ISGR-S). Finally, ISGR-IRR brings additional gains over ISGR-M, including +3.5 on MMT-Bench, +2.4 on RealWorldQA, +2.8 on SEED-Bench, and +122.4 on MME. These results show that interaction-augmented scene graph supervision improves data efficiency, while the IRR stage further strengthens structure-aware reasoning.

4.5 Ablation Study

Effect of Query-Type Design. Figure 5 evaluates the contribution of different query types used in Summarize-and-Align graph construction. We compare the full ISGR setting with variants

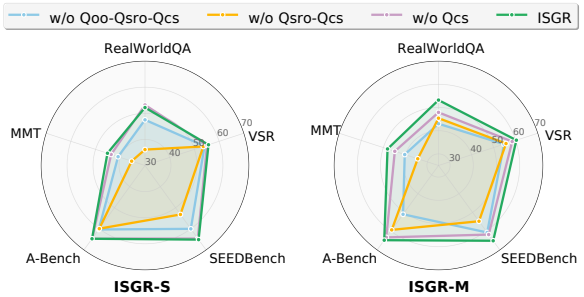

Figure 5: **Ablation study on instruction categories.** Performance comparison of different query type combinations for Summarize-and-Align graph within ISGR (SFT). Highlighted rows indicate the best configuration incorporating all query types.

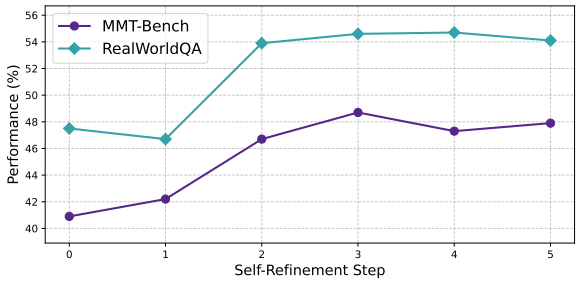

Figure 6: **Impact of Self-Refinement.** Performance improvement across self-refinement iterations on MMT-Bench and RealWorldQA datasets, demonstrating the effectiveness of the progressive data enhancement cycle.

that remove Object-Object queries (Q_{oo}), Subject-Relation-Object queries (Q_{sro}), and Comprehensive Subject queries (Q_{cs}). The full query design achieves the best overall performance, indicating that these query types provide complementary supervision for interaction-aware reasoning. With the 1.3M training set, using all query types obtains an average score of 61.96%, compared with 52.34% when specialized interaction queries are removed. The gain is especially clear on MMT-Bench, where performance improves from 30.3% to 48.1%, showing that query-driven interaction construction is important for complex scene reasoning.

Effect of Long-term Interaction Reinforcement. Table 1 shows that LIR consistently strengthens interaction-aware reasoning. Compared with LLaVA-v1.5-7B-M, ISGR-IRR improves RealWorldQA, MMT-Bench, and SEEDBench by +6.4, +7.6, and +7.2 points, respectively. These gains are mainly observed on benchmarks involving spatial understanding, multi-object interaction, and compositional scene reasoning, indicating that LIR helps the model better internalize role-aware interaction structures.

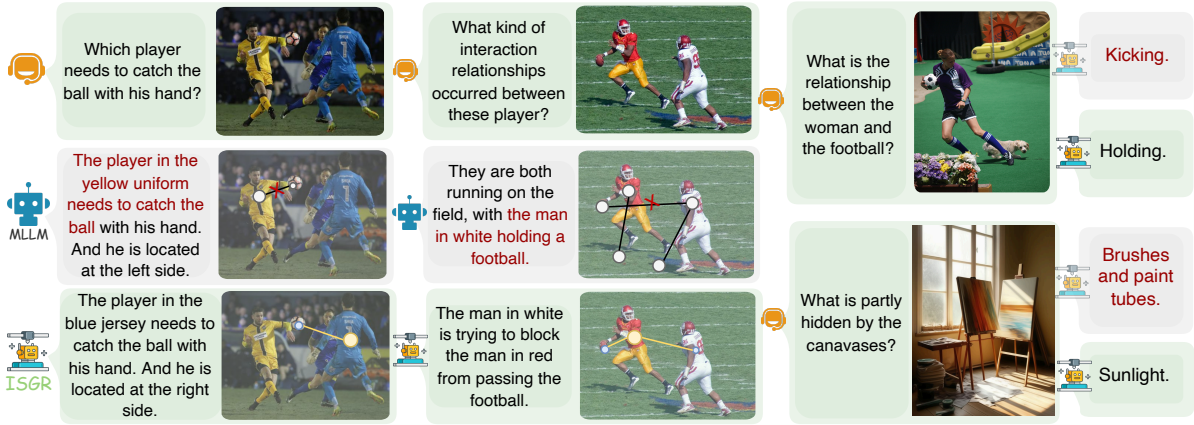

Figure 7: **Case Study.** (Left) Limitations of spatial reasoning where models provide contradictory answers based solely on proximity; (Medium) Scene graph focusing enables accurate identification of functional interactions; (Right) Long-term Interaction Reinforcement enhances subtle relationship identification and generalization to novel scenes, including unseen relationship types in generated images.

Table 2 shows that all reward components are useful. The full reward design obtains the best average performance of 60.97%, confirming the complementarity of Relevance, Disambiguation, and Focus. Disambiguation achieves the strongest individual result of 60.47%, suggesting that reliable subject-object separation is a key factor in complex interaction reasoning.

Impact of Self-Refinement. As shown in Figure 6, self-refinement improves performance in early iterations on both MMT-Bench and RealWorldQA. On MMT-Bench, performance increases from 40.9% to 48.7% after two refinement rounds, while RealWorldQA improves from 47.5% to 54.6%. These gains suggest that the refined interaction graphs provide cleaner role assignments and more focused reasoning cues, thereby improving the quality of interaction-aware training data. However, performance slightly drops in later iterations, indicating that repeated refinement may also accumulate noisy pseudo-labels or over-specialized graph patterns. To avoid such degradation, we select the best checkpoint based on validation performance and apply reward-based graph filtering during refinement.

4.6 Case Study

Figure 7 illustrates how ISGR improves interaction reasoning beyond spatial-only cues. In the first example, the baseline relies mainly on object proximity and incorrectly selects the nearest player. This error shows that spatial closeness alone cannot capture the functional role of each participant. In contrast, ISGR reasons over action roles and scene context, allowing it to identify the player whose behavior and position are most consistent with catching the ball.

The second example shows the benefit of focus-guided interaction modeling. The baseline response drifts toward irrelevant scene details and gives an inaccurate description. By suppressing distracting relations, ISGR identifies the key interaction: "*the man in white is trying to block the man in red from passing the football.*"

The last example highlights the effect of Long-term Interaction Reinforcement. Compared with the SFT-only model, ISGR-IRR better captures the underlying interaction between the entities rather than relying on superficial visual cues. This helps the model identify interaction roles and subtle entity relations more accurately in unseen scenes.

5 Conclusion

We introduced Interaction-augmented Scene Graph Reasoning (ISGR), a framework that enhances vision-language models' reasoning of complex interactions in visual scenes. Our approach comprises three synergistic components: Summarize-and-Align Graph Construction transforms spatial relationships into hierarchical, role-aware interaction graphs; Long-term Interaction Reinforcement combines supervised fine-tuning with structure-aware reinforcement for compositional reasoning; and Progressive Self-Refinement establishes a dynamic feedback loop for continuous improvement. Extensive experiments demonstrate that ISGR significantly outperforms existing methods on interaction-heavy benchmarks while exhibiting strong generalization to novel scenarios.

Limitations

ISGR is currently an offline interaction-reasoning framework rather than a fully end-to-end online framework. Interaction graphs are constructed before training, and model parameters remain fixed during inference. Although this setting enables stable and reproducible evaluation, it does not yet support online graph refinement, continuous feedback integration, or dynamic parameter updating. Moreover, ISGR focuses on static visual scenes and does not explicitly model temporal state changes as in video-based interaction understanding. Future work will explore an online closed-loop extension that can update interaction representations and model behavior as new visual scenes become available.

References

Jinze Bai, Shuai Bai, Shusheng Yang, Shijie Wang, Sinan Tan, Peng Wang, Junyang Lin, Chang Zhou, and Jingren Zhou. 2023. Qwen-vl: A versatile vision-language model for understanding, localization, text reading, and beyond. *Preprint*, arXiv:2308.12966.

Zixiang Chen, Yihe Deng, Huizhuo Yuan, Kaixuan Ji, and Quanquan Gu. 2024. Self-play fine-tuning converts weak language models to strong language models. In *International Conference on Machine Learning*, pages 6621–6642. PMLR.

Wenliang Dai, Junnan Li, Dongxu Li, Anthony Meng Huat Tiong, Junqi Zhao, Weisheng Wang, Boyang Li, Pascale Fung, and Steven Hoi. 2023. Instructblip: Towards general-purpose vision-language models with instruction tuning. *Preprint*, arXiv:2305.06500.

Bruno Cesar de Oliveira Souza, Marius Aasan, Helio Pedrini, and Adin Ramirez Rivera. 2023. Self-graphvqa: a self-supervised graph neural network for scene-based question answering. In *Proceedings of the IEEE/CVF international conference on computer vision*, pages 4640–4645.

Penghui Du, Yu Wang, Yifan Sun, Luting Wang, Yue Liao, Gang Zhang, Errui Ding, Yan Wang, Jingdong Wang, and Si Liu. 2024. Lami-detr: Open-vocabulary detection with language model instruction. In *European Conference on Computer Vision*, pages 312–328. Springer.

Haodong Duan, Junming Yang, Yuxuan Qiao, Xinyu Fang, Lin Chen, Yuan Liu, Xiaoyi Dong, Yuhang Zang, Pan Zhang, Jiaqi Wang, and 1 others. 2024. Vlmevalkit: An open-source toolkit for evaluating large multi-modality models. In *Proceedings of the 32nd ACM international conference on multimedia*, pages 11198–11201.

Mingtao Feng, Chenbo Yan, Zijie Wu, Weisheng Dong, Yaonan Wang, and Ajmal Mian. 2025. History-enhanced 3d scene graph reasoning from rgb-d sequences. *IEEE Transactions on Circuits and Systems for Video Technology*, 35(8):7667–7682.

Elias Greve, Martin Büchner, Niclas Vödisch, Wolfram Burgard, and Abhinav Valada. 2024. Collaborative dynamic 3d scene graphs for automated driving. In *2024 IEEE International Conference on Robotics and Automation (ICRA)*, pages 11118–11124. IEEE.

Danna Gurari, Qing Li, Abigale J Stangl, Anhong Guo, Chi Lin, Kristen Grauman, Jiebo Luo, and Jeffrey P Bigham. 2018. Vizwiz grand challenge: Answering visual questions from blind people. In *Proceedings of the IEEE conference on computer vision and pattern recognition*, pages 3608–3617.

Yufeng Huang, Jiji Tang, Zhuo Chen, Rongsheng Zhang, Xinfeng Zhang, Weijie Chen, Zeng Zhao, Zhou Zhao, Tangjie Lv, Zhipeng Hu, and 1 others. 2024. Structure-clip: Towards scene graph knowledge to enhance multi-modal structured representations. In *Proceedings of the AAAI conference on artificial intelligence*, volume 38, pages 2417–2425.

Alina Kuznetsova, Hassan Rom, Neil Alldrin, Jasper Uijlings, Ivan Krasin, Jordi Pont-Tuset, Shahab Kamali, Stefan Popov, Matteo Malloci, Alexander Kolesnikov, and 1 others. 2020. The open images dataset v4: Unified image classification, object detection, and visual relationship detection at scale. *International journal of computer vision*, 128(7):1956–1981.

Bo Li, Yuanhan Zhang, Dong Guo, Renrui Zhang, Feng Li, Hao Zhang, Kaichen Zhang, Peiyuan Zhang, Yanwei Li, Ziwei Liu, and 1 others. 2024a. Llava-onevision: Easy visual task transfer. *arXiv preprint arXiv:2408.03326*.

Bohao Li, Yuying Ge, Yixiao Ge, Guangzhi Wang, Rui Wang, Ruimao Zhang, and Ying Shan. 2024b. Seed-bench: Benchmarking multimodal large language models. In *Proceedings of the IEEE/CVF Conference on Computer Vision and Pattern Recognition*, pages 13299–13308.

Bin Lin, Zhenyu Tang, Yang Ye, Jiaxi Cui, Bin Zhu, Peng Jin, Jinfa Huang, Junwu Zhang, Yatian Pang, Munan Ning, and 1 others. 2024. Moe-llava: Mixture of experts for large vision-language models. *arXiv preprint arXiv:2401.15947*.

Fangyu Liu, Guy Emerson, and Nigel Collier. 2023a. Visual spatial reasoning. *Transactions of the Association for Computational Linguistics*, 11:635–651.

Fuxiao Liu, Kevin Lin, Linjie Li, Jianfeng Wang, Yaser Yacoob, and Lijuan Wang. 2023b. Mitigating hallucination in large multi-modal models via robust instruction tuning. *arXiv preprint arXiv:2306.14565*.

Haotian Liu, Chunyuan Li, Yuheng Li, and Yong Jae Lee. 2024. Improved baselines with visual instruction tuning. In *Proceedings of the IEEE/CVF Conference on Computer Vision and Pattern Recognition*, pages 26296–26306.

Haotian Liu, Chunyuan Li, Qingyang Wu, and Yong Jae Lee. 2023c. Visual instruction tuning. *Advances in neural information processing systems*, 36:34892–34916.

DataCanvas Ltd. 2024. mmalaya. https://github.com/DataCanvasIO/MMAlaya.

Vladislav Makarov, Mark Gizetdinov, and Dmitry Yudin. 2026. Scenegraphvlm: Dynamic scene graph generation from video with vision-language models. *arXiv preprint arXiv:2605.13667*.

Andrés Marafioti, Orr Zohar, Miquel Farré, Merve Noyan, Elie Bakouch, Pedro Cuenca, Cyril Zakka, Loubna Ben Allal, Anton Lozhkov, Nouamane Tazi, and 1 others. 2025. Smolvlm: Redefining small and efficient multimodal models. *arXiv preprint arXiv:2504.05299*.

Zhiliang Peng, Wenhui Wang, Li Dong, Yaru Hao, Shaohan Huang, Shuming Ma, Qixiang Ye, and Furu Wei. 2024. Grounding multimodal large language models to the world. In *International Conference on Learning Representations*, volume 2024, pages 51575–51598.

Zhihong Shao, Peiyi Wang, Qihao Zhu, Runxin Xu, Junxiao Song, Xiao Bi, Haowei Zhang, Mingchuan Zhang, YK Li, Y Wu, and 1 others. 2024. Deepseekmath: Pushing the limits of mathematical reasoning in open language models. *arXiv preprint arXiv:2402.03300*.

Jingyi Wang, Jianzhong Ju, Jian Luan, and Zhidong Deng. 2025a. Llava-sg: Leveraging scene graphs as visual semantic expression in vision-language models. In *ICASSP 2025-2025 IEEE International Conference on Acoustics, Speech and Signal Processing (ICASSP)*, pages 1–5. IEEE.

Junke Wang, Lingchen Meng, Zejia Weng, Bo He, Zuxuan Wu, and Yu-Gang Jiang. 2023. To see is to believe: Prompting gpt-4v for better visual instruction tuning. *arXiv preprint arXiv:2311.07574*.

Weizhen Wang, Chenda Duan, Zhenghao Peng, Yuxin Liu, and Bolei Zhou. 2025b. Embodied scene understanding for vision language models via metavqa. In *Proceedings of the Computer Vision and Pattern Recognition Conference*, pages 22453–22464.

Xinlong Wang, Xiaosong Zhang, Zhengxiong Luo, Quan Sun, Yufeng Cui, Jinsheng Wang, Fan Zhang, Yueze Wang, Zhen Li, Qiying Yu, and 1 others. 2024. Emu3: Next-token prediction is all you need. *arXiv preprint arXiv:2409.18869*.

Shengqiong Wu, Hao Fei, Jingkang Yang, Xiangtai Li, Juncheng Li, Hanwang Zhang, and Tat-seng Chua. 2025a. Learning 4d panoptic scene graph generation from rich 2d visual scene. In *Proceedings of the Computer Vision and Pattern Recognition Conference*, pages 24539–24549.

Shun-Cheng Wu, Johanna Wald, Keisuke Tateno, Nassir Navab, and Federico Tombari. 2021. Scenegraphfusion: Incremental 3d scene graph prediction from rgb-d sequences. In *Proceedings of the IEEE/CVF Conference on Computer Vision and Pattern Recognition*, pages 7515–7525.

Tianhao Wu, Weizhe Yuan, Olga Golovneva, Jing Xu, Yuandong Tian, Jiantao Jiao, Jason E Weston, and Sainbayar Sukhbaatar. 2025b. Meta-rewarding language models: Self-improving alignment with llm-as-a-meta-judge. In *Proceedings of the 2025 Conference on Empirical Methods in Natural Language Processing*, pages 11548–11565.

x.ai. 2024. Grok-1.5 vision preview. Accessed: 2025-03-23.

Linhui Xiao, Xiaoshan Yang, Xiangyuan Lan, Yaowei Wang, and Changsheng Xu. 2025. Towards visual grounding: A survey. *IEEE Transactions on Pattern Analysis and Machine Intelligence*, 1(1):1–20.

Shuo Yang, Caren Han, Siwen Luo, and Eduard Hovy. 2025. Magic-vqa: Multimodal and grounded inference with commonsense knowledge for visual question answering. In *Findings of the Association for Computational Linguistics: ACL 2025*, pages 16967–16986.

Kaining Ying, Fanqing Meng, Jin Wang, Zhiqian Li, Han Lin, Yue Yang, Hao Zhang, Wenbo Zhang, Yuqi Lin, Shuo Liu, and 1 others. 2024. Mmt-bench: A comprehensive multimodal benchmark for evaluating large vision-language models towards multitask agi. In *International Conference on Machine Learning*, pages 57116–57198. PMLR.

Haobo Yuan, Xiangtai Li, Chong Zhou, Yining Li, Kai Chen, and Chen Change Loy. 2024a. Open-vocabulary sam: Segment and recognize twenty-thousand classes interactively. In *European Conference on Computer Vision*, pages 419–437. Springer.

Weizhe Yuan, Richard Yuanzhe Pang, Kyunghyun Cho, Xian Li, Sainbayar Sukhbaatar, Jing Xu, and Jason Weston. 2024b. Self-rewarding language models. In *Proceedings of the 41st International Conference on Machine Learning*, pages 57905–57923.

Wenzhe Zhai, Xianglei Xing, Mingliang Gao, and Qilei Li. 2025. Zero-shot object counting with vision-language prior guidance network. *IEEE Transactions on Circuits and Systems for Video Technology*, 35(3):2487–2498.

Zicheng Zhang, Haoning Wu, Chunyi Li, Yingjie Zhou, Wei Sun, Xiongkuo Min, Zijian Chen, Xiaohong Liu, Weisi Lin, and Guangtao Zhai. 2024. A-bench:

Are lmms masters at evaluating ai-generated images? *arXiv preprint arXiv:2406.03070*.

Changmeng Zheng, Dayong Liang, Wengyu Zhang, Xiao-Yong Wei, Tat-Seng Chua, and Qing Li. 2024. A picture is worth a graph: A blueprint debate paradigm for multimodal reasoning. In *Proceedings of the 32nd ACM International Conference on Multimedia*, pages 419–428.

Wenlve Zhou and Zhiheng Zhou. 2024. Unsupervised domain adaption harnessing vision-language pre-training. *IEEE Transactions on Circuits and Systems for Video Technology*, 34(9):8201–8214.

Yifeng Zhu, Jonathan Tremblay, Stan Birchfield, and Yuke Zhu. 2021. Hierarchical planning for long-horizon manipulation with geometric and symbolic scene graphs. In *2021 IEEE International Conference on Robotics and Automation (ICRA)*, pages 6541–6548. IEEE.

A More Experimental Details

A.1 Algorithm for ISGR

For a better understanding of ISGR, an algorithmic procedure has been formulated in Algorithm 1.

Algorithm 1 ISGR (Interaction-augmented Scene Graph Reasoning)

Require: Input S = (question Q, Wild Image I_{max}), Multimodal LLM $\mathcal{M}$, maximum iterations T

for $t = 1$ to T **do**
  // Phase 1: Hierarchical Graph Construction
  $\mathcal{G}^0 \leftarrow f_{\text{init}}(I, Q)$
  $\mathcal{G}^s \leftarrow f_{\text{spatial}}(I, Q, \mathcal{M})$
  $\mathcal{G}^a \leftarrow f_{\text{abstract}}(\mathcal{G}^s, I, Q)$
  $\mathcal{G}^t \leftarrow f_{\text{ICoT}}(\mathcal{G}^a, I, Q)$
  $\mathcal{G}^{\text{final}} \leftarrow f_{\text{abstract}}(\mathcal{G}^t, I, Q)$
end for
// Phase 2: Dual-Phase Learning
$\mathcal{L}_{\text{SFT}} \leftarrow \mathbb{E}_{(I,Q,\mathcal{G}^{\text{final}})}\left[-\log P_{\mathcal{M}}(\mathcal{G}^{\text{final}}|I, Q)\right]$
$\mathcal{L}_{\text{IRR}} \leftarrow \mathbb{E}_{(I,Q,A)\sim\mathcal{D}}\left[R(A_{\text{pred}}, A_{\text{gt}})\right]$
$\mathcal{M}_t^{\text{enh}} \leftarrow \text{GRPO}(\mathcal{M}, \mathcal{D}_t^{\text{inter}}, R)$
$\mathcal{A} \leftarrow P_{\mathcal{M}_T^{\text{enh}}}(\cdot \mid I, Q, \mathcal{G}_T^{\text{final}})$
return $\mathcal{A}, \mathcal{M}_T^{\text{enh}}$

A.2 Statistics of Datasets

The training dataset and source code will be accessed. Table 3 summarizes the key statistics of the constructed dataset designed for interaction-focused scene graph learning. This dataset integrates multiple sources, including LLaVA-v1.5-mixed-665k(Liu et al., 2024), LVIS-Instruct-4V(Wang et al., 2023), LRV-Instruct(Liu et al., 2023b), and OpenImages(Kuznetsova et al., 2020), totaling 1,371K instances in the medium-scale variant. The small-scale variant includes 841K instances, combining LLaVA with interaction-augmented data. Additionally, we have a curated set of 1k high-quality instructions specifically designed for Interaction Reasoning Reinforcement.

Dataset	Source	Size
LLaVA-v1.5-mixed	LLaVA	665K
LVIS-Instruct-4V	LVIS	300K
LRV-Instruct	LRV	300K
OpenImages	OpenImages	176K
Interaction-Augmented	OpenImages+	300K
Small-Scale Variant	LLaVA + Aug	841K
Medium-Scale Variant	Whole	1,371K
IRR Instruction	Curated Set	1k

Table 3: Statistics of the constructed dataset for interaction-focused scene graph learning. Each variant includes interaction-augmented scene graph data to support robust reasoning capabilities.

Setting	Value
Language Model (LLM)	Vicuna-7B
Vision Encoder	CLIP-L/14
Hardware Requirement	8x A100 (40GB)
Truncation Mode	Left
Number of Beams	1
Batch Size	8
Temperature	0.2
Top-p	0.9
Data Type	float16
Image Resolution	224x224
Maximum Input Length	512
Maximum Output Length	128
Train Time for ISGR-S	13 hours
Train Time for ISGR-M	25 hours
Train Time for ISGR-IRR	34 min
Inference Time for VQAv2	3.87 sample/s
Inference Time for SEEDBench	3.52 sample/s
Inference Time for MMT-Bench	3.57 sample/s

Table 4: ISGR Model Fine-Tuning and Inference Settings

A.3 Model Deployment

The specifics of model deployment and hyperparameter configurations for the ISGR model are detailed in Table 4, highlighting hardware requirements, training parameters, and inference times across various benchmarks.

A.4 Prompts

A.4.1 Spatial initialization

You are an AI assistant. Generate a spatial scene graph identifying objects and their spatial relationships in the given image. Use the format of relationship triples: **[subject, relation, object]**
Input: **{image}**
Output: **{spatial scene graph}**

A.4.2 Abstract graph

You are a vision-language assistant. Given the spatial scene graph and image, produce an abstract graph by: 1. Filtering background and irrelevant elements;
2. Retaining essential objects and spatial relationships that define the core scene activity;
3. Ensuring all entities are unambiguous.
Input: **{spatial scene graph}, {image}**
Output: **{abstract graph}**

A.4.3 Interaction Knowledge

You are an AI assistant. Using the *abstract graph*, identify all interactions between entities that are relevant to the core scene depicted in the image.
Your task: 1. Analyze the abstract graph to extract meaningful interactions.
2. For each interaction, specify the subject, action, and object.
3. Ensure that all entities are clearly defined and unambiguous.
Input: **{abstract graph}, {image}**
Output: **{interaction knowledge}**

A.4.4 Interaction Graph

You are an AI assistant. Using the *interaction knowledge*, further abstract the scene by identifying the most relevant interactions for the core scene depicted in the image.
Your task: 1. Focus on the essential interactions that define the dynamics of the scene.
2. For each interaction, specify the subject, action, and object.
3. Ensure clarity by adhering to the saliency, grounding, and consistency constraints.
Input: **{interaction knowledge}, {image}**
Output: **{interaction graph}**

A.4.5 Querying the Interaction-augmented Graph

You are a vision-language assistant. Given the scene graph below, generate **4 QA pairs** covering the following question types:
Q1: What is the relationship between `obj1[bbox]` and `obj2[bbox]`? **A**: `obj1 rel obj2`.
Q2: What does `obj1[bbox]` do? **A**: `obj1 rel obj2[bbox]`.
Q3: What is `rel` by `obj2[bbox]`? **A**: `obj1[bbox] rel obj2`.
Q4: What relationships does `obj1[bbox]` have? **A**: `obj1 rel1 obj2[bbox]`, `rel2 obj3[bbox]`
Rules: use exact object and relation names from the graph; include bbox coordinates in questions for identification, but only in answers for objects not already specified in the question; only generate answerable QA pairs from the provided graph.
Input: **{interaction graph}, {image}**
Output: **{interaction instruction}**

B Details of Reinforcement Learning Framework

The reinforcement stage takes the SFT model, an image-question pair (I, Q), and its interaction graph G as input. The graph contains role-aware triplets $T(G) = \{(s, a, o)\}$, where s, a, and o denote the subject, action, and object. For each input, the policy samples K responses $\{y_k\}_{k=1}^{K}$. We parse each response into predicted triplets $\hat{T}_k$ and mentioned entities $\hat{E}_k$, and compare them with the graph-supported evidence.

The reward is computed as:

$$R(y_k) = 0.4R_{\text{rel}}(y_k) + 0.4R_{\text{dis}}(y_k) + 0.2R_{\text{foc}}(y_k). \tag{11}$$

The three terms are rule-based scores derived from the constructed graph:

$$\begin{aligned} R_{\text{rel}} &= \frac{|\hat{T}_k \cap T(G)|}{\max(|\hat{T}_k|, 1)}, \\ R_{\text{dis}} &= \frac{1}{\max(|\hat{T}_k|, 1)} \sum_{\hat{t} \in \hat{T}_k} \mathbb{I}[\text{role}(\hat{t}) = \text{role}(t^*)], \\ R_{\text{foc}} &= \frac{|\hat{E}_k \cap E(G_Q)|}{\max(|\hat{E}_k|, 1)}. \end{aligned} \tag{12}$$

Here, t^* is the best-matched triplet in G, and G_Q is the query-relevant subgraph. R_{rel} checks whether the response is supported by graph triplets, R_{dis}

checks whether subject-object roles are correctly assigned, and R_{foc} checks whether the response focuses on question-relevant entities. All rewards are computed by deterministic matching; no learned reward model is used.

We optimize the model with GRPO. For each sampled group, the advantage is normalized within the group:

$$\hat{A}_k = \frac{R(y_k) - \text{mean}(R)}{\text{std}(R) + \epsilon},$$
$$\mathcal{L}_{\text{IRR}} = -\frac{1}{K}\sum_{k=1}^{K} \hat{A}_k \log \pi_\theta(y_k|I, Q). \tag{13}$$

The output is the interaction-reinforced model. To improve stability, invalid responses are filtered before reward computation, rewards are normalized within each group, and the RL stage is kept short to avoid overfitting to noisy graph signals.

C Details of Progressive Self-Refinement

The self-refinement stage takes the current model M_t and dataset $D_t = \{(I_i, Q_i, G_i^t)\}_{i=1}^N$ as input. For each (I_i, Q_i), the model regenerates a candidate interaction graph:

$$\hat{G}_i^{t+1} = f_{\text{ICoT}}(I_i, Q_i, M_t). \tag{14}$$

The candidate graph is accepted only if it passes graph-level filtering. We compute:

$$S(\hat{G}) = \alpha S_{\text{valid}} + \beta S_{\text{rel}} + \gamma S_{\text{cons}}, \tag{15}$$

where S_{valid} measures whether triplets are complete, S_{rel} measures whether entities are relevant to the question, and S_{cons} measures whether the new graph is consistent with the previously accepted graph.

The update rule is:

$$G_i^{t+1} = \begin{cases} \hat{G}_i^{t+1}, & S(\hat{G}_i^{t+1}) > \tau, \\ G_i^t, & \text{otherwise.} \end{cases} \tag{16}$$

The output is the refined dataset $D_{t+1} = \{(I_i, Q_i, G_i^{t+1})\}_{i=1}^N$. To reduce error accumulation, ISGR uses conservative replacement: low-score graphs are rejected, original instruction data are retained as anchors, duplicated graph patterns are removed, and refinement stops when validation performance no longer improves.

Model	MMT	A-Bench	SEEDBench
ISGR	45.2	64.7	65.1
-SG/Bbox	43.7(-1.5)	63.1(-1.6)	62.9(-2.2)
ISGR-IRR	48.7	66.3	67.9
-SG/Bbox	47.8(-0.9)	65.1(-1.2)	66.4(-1.5)

Table 5: Impact of spatial grounding on model performance. While spatial grounding consistently improves results, LIR shows robust reasoning capabilities even without explicit spatial information.

C.1 More Ablation Study

Impact of Spatial Grounding. To investigate the contribution of grounding bounding boxes in scene graph understanding, we conducted ablation experiments by removing bounding box information from both our Long-term Interaction Reinforcement(LIR) training processes. Table 5 presents the comparative results across multiple benchmarks.

The experimental results demonstrate that grounding information consistently improves model performance across all evaluated datasets. For the base ISGR, removing bounding box information leads to performance drops ranging from 0.5% to 2.2%. This decline is particularly noticeable on specialized visual reasoning benchmarks like MMT-Bench (-1.5%) and SEEDBench (-2.2%), suggesting that spatial grounding plays a crucial role in complex visual understanding tasks.

For the base ISGR, removing bounding boxes leads to drops ranging from 1.5 to 2.2 points. For ISGR-IRR, the drops are 0.9, 1.2, and 1.5 points on MMT-Bench, A-Bench, and SEEDBench, respectively. This highlights the importance of spatial grounding for scene interaction reasoning. However, it is notable that ISGR-IRR without bounding boxes still outperforms the fully-equipped ISGR model on most benchmarks, demonstrating that LIR can elicit strong reasoning capabilities even without explicit spatial grounding.

These findings underscore the significance of spatial grounding in visual reasoning tasks while revealing that our SFT+IRR approach can stimulate powerful reasoning abilities even in its absence. Nevertheless, when spatial grounding is provided within the LIR, it further enhances the model's capabilities in scene interaction reasoning, suggesting that the combination of LIR and explicit spatial information yields the most robust visual understanding performance.

Scene Graph Quality Assessment. To evaluate

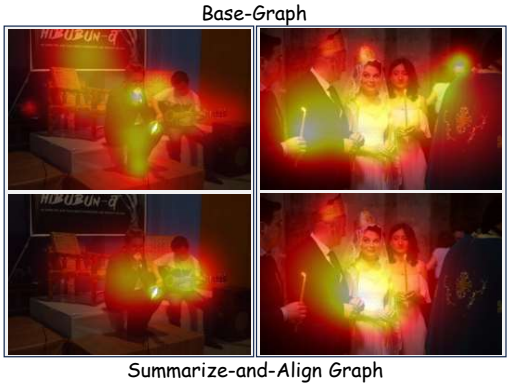

Figure 8: Scene Graph Relationship-based Heatmap Visualization. The heatmap is generated based on scene graph relationships, where regions with more relationships exhibit higher heat intensity

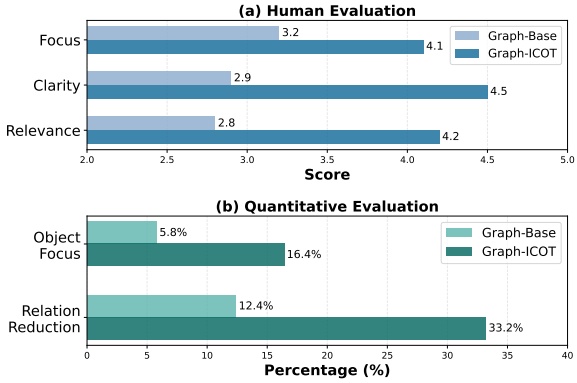

Figure 9: **Evaluation of graph quality.** (a) Human evaluation on Drift, Disambiguation, and Role using a 1-5 scale. (b) Quantitative evaluation on Object Focus and Relation Reduction percentages. Graph-ICoT consistently outperforms Graph-Base across all metrics.

the quality of our generated scene graphs, we conducted a comprehensive human evaluation study. Five evaluators assessed 100 pairs of graphs (original vs. optimized) based on three key criteria: Role, disambiguation, and Drift, using a scale of 1 to 5.

Figure 9(a) presents the comparative results, showing that optimized graphs received significantly higher scores across all metrics. The optimized graphs achieved role scores of 4.2 compared to 2.8 for baseline graphs, disambiguation scores of 4.5 vs. 2.9, and drift scores of 4.1 vs. 3.2. These results demonstrate the effectiveness of our ICOT processing in producing higher-quality scene representations.

Graph Construction Optimization Analysis. We quantified the effectiveness of our scene graph optimization through detailed analysis of object focus and relationship pruning. Figure 9(b) shows the improvements achieved through our Abstract Graph Construction and ICOT processes.

Method	RWQA	MMT	ABench	SEEDB
LLAVA(w/IRR)	51.7	43.2	63.4	63.5
ISGR(wo/ICOT)	52.8	46.4	64.9	65.8
ISGR(w/ICOT)	**54.6**	**48.7**	**66.3**	**67.9**

Table 6: Component ablation study results. Our complete framework with both ICOT and IRR achieves the best performance across all benchmarks.

The results demonstrate substantial improvements in both relation reduction and object focus. In the Abstract Graph Construction phase, we observed a 12.4% improvement in relation reduction and a 5.8% increase in object focus. After applying ICOT processing, these improvements increased dramatically to 33.2% for relation reduction and 16.4% for object focus, highlighting the effectiveness of our interaction-focused optimization approach.

The enhanced focusing capability is further illustrated through relationship-based heatmap visualization (Figure 8), where heat intensity corresponds to relationship density. The results show that our Summarize-and-Align framework significantly improves scene focusing ability, benefiting from effective pruning of complex scene relationships while preserving semantically critical connections.

Component Ablation Study. To validate the individual contributions of our framework components, we conducted comprehensive ablation experiments. Table 6 presents the performance comparison across multiple benchmarks when different components are removed or modified.

The results demonstrate that each component of our framework plays a crucial role. Using only IRR without ICOT's scene graph (LLAVA w/IRR) results in reduced performance compared to our full method. The inclusion of ICOT without IRR (ISGR wo/ICOT) shows improvement over the baseline but still falls short of the full system. Our complete framework (ISGR w/ICOT) achieves the best performance across all benchmarks, with an average improvement of 3.93% over LLAVA w/IRR and 1.91% over ISGR wo/ICOT.

Inference Time. To address concerns about the computational overhead of our approach, we provide a detailed analysis of inference times across different benchmarks. It is important to note that the two LLM calls in our ISGR approach occur during offline training and do not add overhead during

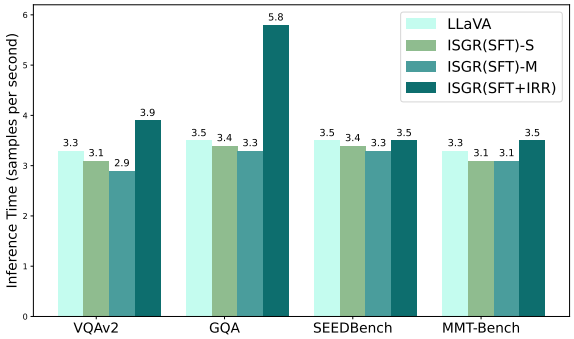

Figure 10: Inference Time

validation. As demonstrated in Figure 10, there are no significant changes in testing times between the model versions.

The results show that our ISGR models maintain competitive inference speeds, with the ISGR-IRR model even achieving faster performance on the SEEDBench benchmark (3.52 samples/s compared to 2.79 samples/s for the baseline). This demonstrates that our interaction-augmented scene graph reasoning does not introduce significant computational overhead during inference, making it practical for real-world applications.

C.2 Fail Case Study

We also encountered limitations within our framework. For instance, as illustrated in the Figure 11, there are scenarios where ISGR does not perform optimally.

The first two failure cases focus on single objects that require the recognition of subtle emotional changes. These scenarios necessitate a high sensitivity to minute facial expressions, which our framework struggles to capture due to its design focus on complex interactions rather than individual emotional nuances.

The third failure case involves multiple objects but pertains to reasoning about physical shapes. In this instance, the task requires making judgments based on geometric characteristics. The ISGR framework, which emphasizes interaction and relationship reasoning, does not adequately address the geometric properties of individual objects, leading to inaccuracies in responses.

These limitations highlight the need for further optimization of the ISGR framework in tasks that require detailed emotional analysis and geometric reasoning. Future research should explore enhancements to better accommodate these specific recognition tasks, thereby expanding the effectiveness of our approach across a broader range of application scenarios.

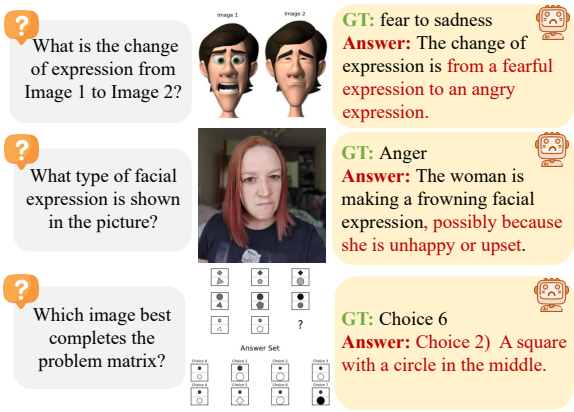

Figure 11: **Failure cases.** Failure cases illustrating the limitations of our framework. The first two cases highlight challenges in recognizing subtle emotional changes in single objects, while the third involves reasoning about physical shapes among multiple objects. These scenarios reveal the constraints of our ISGR framework in nuanced emotional recognition and geometric reasoning.